日本語発話時の口形変化量の分析と発話映像自動生成への適用

宮崎 剛[1]　中島 豊四郎[2]

概要：著者らは，これまで機械読唇に関する研究を進めてきた．その中で，日本語語句の仮名表記からその語句を発話する際の発話映像を再現する方法を提案した．この方法では，まず，仮名表記から口形順序コードと呼ぶ口形記号の順列を生成する．口形順序コードは，語句発話時に形成される口形の順を表現している．発話映像の生成は，この口形順序コードに対応する口形を表示させながら，CG 技術のモーフィングを利用して口形間の口形変化を実現する．ただし，これまでの方法では，口形の変化量を実際の発話映像を参考にしながら実験的に決定していた．そのため，口形の変化に違和感のある映像が生成される場合もあった．そこで本論文では，高いフレームレートで撮影した実際の発話映像から口形の変化量を分析し，その結果に基づいて発話映像を生成する方法を提案する．そして，生成された映像について，被験者による主観評価の結果を示し，提案手法の有効性を評価する．

Analysis of Mouth Shape Deformation Rate and its Application to Automatic Generation of Japanese Utterance Images

Tsuyoshi Miyazaki[1]　Toyoshiro Nakashima[2]

1. はじめに

著者らはこれまで，聴覚障害者のコミュニケーションを支援する技術に関する研究を進めてきた．聴覚障害者が音声以外の情報を使ってコミュニケーションを取る方法には，主に “手話” や “筆談”，“読唇” が挙げられる．手話は有効な手段であるが，コミュニケーションを取る相手も手話を習得していなければならないため，その相手は限られてしまう．筆談は，多くの相手とコミュニケーションを取ることができるが，意志の疎通に時間がかかるため，会話の手段としてはあまり有効ではない．読唇も筆談と同様，多くの相手とコミュニケーションを取ることができるが，聞き手に読唇の能力が必要となる．ただし，読唇の場合は，手話のように会話の相手にその技能を要求するのではなく，聞き手（本人）側にその能力が必要とされる．読唇の能力の習得は困難であるが，この能力を習得できれば多くの人と円滑なコミュニケーションを取ることができるようになるため，有効な手段であると言える．

現在，読唇技能の習得には，人と人との対面による学習と，予め撮影されたビデオ映像を利用した学習がある．前者は学習効果が高いが，いつでも学習できるわけではなく，話し側もある程度読唇を教えられる能力を持っていなければならない．後者は学習者の都合で学習できるが，任意の語句を話す口唇の動きには対応できない．そのため，ビデオ映像を用いた学習はその習熟度に個人差が出やすい．これらのことから，任意の語句を話す発話映像が生成できれば，読唇の教材として利用可能ではないかと考える．

コンピュータを用いて発話映像を生成する研究として，発話動画像合成（Visual Speech Synthesis）がある．これは，発話の音声と同期した口唇の動画像を生成する研究で，発話音声からパラメータを生成するなどして CG の 3 次元顔モデルに対し，口唇や顔形状の動きを合成する方法である [1], [2]．また，擬人化音声対話エージェントに関する研

[1] 神奈川工科大学情報学部情報工学科
Department of Information and Computer Sciences, Kanagawa Institute of Technology, Atsugi, Kanagawa 243–0292, Japan
[2] 椙山女学園大学文化情報学部文化情報学科
School of Culture-Information Studies, Sugiyama Jogakuen University, Nagoya, Aichi 464–8662, Japan

究では，CG モデルのエージェントが合成音声と同期して口唇を動作させ，対話をするような振る舞いを示す方法が提案されている [3]．これらの研究では音声と口唇動作との同期を目指しているが，本研究は聴覚障害者が読唇技術を習得する教材としての発話映像を目指している．そのため，本研究では音声を利用せずに，正確な口唇動作を表現するの動画像の生成を目指している．さらに，発話動画像合成では，口形変化を決定する方法として隠れマルコフモデル（HMM）などの統計的なデータを用いているが [2]，日本語には特有の音と口形の関係が存在するため [4]，本研究ではこの関連性を活かしたルールベースの手法を用いている．

これまでに著者らは，日本語語句の仮名表記からその語句を発話する際の無声発話映像を自動生成する方法を提案した [5]．これにより，任意の語句を正確な口形変化で表現できる発話映像を生成することが可能となった．そして，この機能を Android 端末で動作するアプリケーションとして試作し，評価を行った [6]．しかし，これまでの映像では映像のフレームレートが 30fps であったことと，口形の変形率が著者の観測から設定したものであったため，発話映像に違和感があった．

そこで本論文では，高いフレームレートで撮影できるカメラを用いて発話映像を撮影し，口唇の詳細な動作を分析し，発話時の口形変化率を導出する．そして，得られた口形変化率を基に発話映像を生成し，被験者に主観評価をしてもらい，発話映像を評価する．

2. 口形変化量の分析

日本語発話時の両唇の動きを分析するため，1 秒間あたり 150 フレームの画像を取得できる，高いフレームレートのカメラを用いて発話時の口唇周辺を撮影した．その際，口唇の動作を分析しやすくするため，**図 1** に示すように，上唇と下唇の中央付近と両口角の合計 4 箇所に青いマーカを貼り付けた．また，発話時に顔が上下左右に動くと両唇の動きの分析に影響が出るため，鼻の領域を検出して簡易的な口唇領域のトラッキングを行うことで，発話時の顔の動きにも考慮した．

2.1 マーカの時系列座標取得

取得した発話映像からマーカを検出する処理を**図 2** に示す．カメラで取得したフレームは RGB 色空間の画像であるため，最初にこれを HSV 色空間へ変換する．そして，変換した画像から色相（Hue）のプレーンのみを取り出し，マーカの青い色（以降，マーカ色と表現する）の範囲を抽出して二値化する．このとき，ノイズによる影響で，マーカ領域以外の画素でもその色がマーカ色の色相範囲内にあるとその部分も検出されてしまうため，この領域を取り除かなければならない．また，撮影時の光の影響やカメラの特性等により，マーカの領域内であっても画素の色がマーカ色の色相範囲外になると抽出されない領域が出てくる．ただし，これらの画素はまとまった大きな領域ではなく，点在する小さな領域であることがほとんどであるため，オープニングやクロージングと呼ばれる画像処理を施すことで，これらの領域を取り除くことができる．そして，最後に残ったある一定以上の大きさ（面積）を持つ領域に対して輪郭の検出を行い，そのサイズから 4 つのマーカの領域を検出し，その中心座標を取得する．これらの処理をすべてのフレームに対して順に行い，4 つのマーカの時系列の座標を取得する．

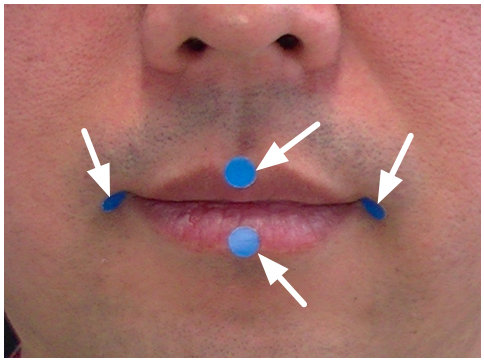

図 1 口唇動作分析のために付けた青いマーカ（4 箇所）
Fig. 1 Four markers to analyze the movement of lips.

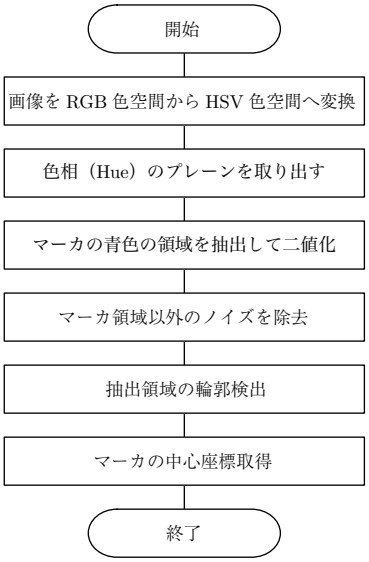

図 2 マーカを検出する流れ
Fig. 2 Marker detection flow.

例として，口唇領域画像からマーカ色の領域を検出し，オープニングとクロージングでノイズ除去する過程を**図 3** に示す．図 3 右上の画像は，同図左上の画像からマーカ色領域を抽出し，二値化したものである．白の領域が対象とするマーカの領域で，黒の領域が背景である．この画像では，上唇に貼り付けたマーカ領域内に小さな抜け（黒い領域）と，歯と口内の境界部分に 6 箇所のごく小さな誤検出

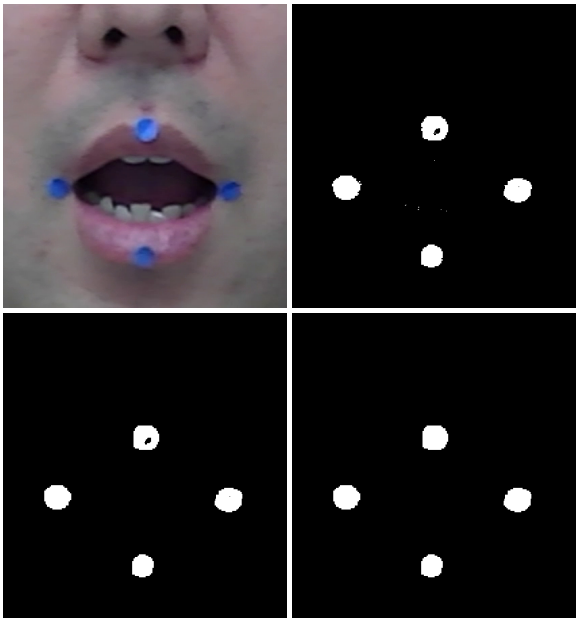

図 3　口唇領域画像（左上）と色相プレーンでマーカ色を抽出した二値画像（右上），オープニング画像（左下），クロージング画像（右下）

Fig. 3　Captured mouth image (top left), Extracted and binarized the marker area (top right), The image after Opening process (bottom left) and The image after Closing process (bottom right).

領域（白い領域）が確認できる．そこで，オープニング処理を施すことで誤検出の領域が取り除かれ（図 3 左下），クロージング処理を施すことでマーカ領域内の小さな抜けが取り除かれる（図 3 右下）．

2.2　口形変化の近似式

得られた 4 つのマーカの時系列の座標変化を分析した結果，上唇と左右の口角は大きく変化しないことが判明した．また，下唇についても x 座標の値はほとんど変化せず，y 座標の値のみが大きく変化していた．**図 4** に，“まみむめも” と発話した際の下唇の y 座標の時系列変化のグラフを示す．

図 4 のグラフでは，座標値が山型に変化する部分で下唇が動いていることになり，右上がりに変化している部分では下唇が下方向に，右下がりに変化している部分では下唇が上方向に動いていることになる．そこで，図 4 中の破線で示した区間（192 フレームから 208 フレーム）を取り出したところ，**図 5** のように 3 次式で近似できることが判明した．

3.　発話映像生成方法

発話時の唇の動きが 3 次式で表現できることが判明したため，これを用いて口形変形画像を生成する．発話映像の生成には，著者らがこれまでに提案してきた方法を利用する．

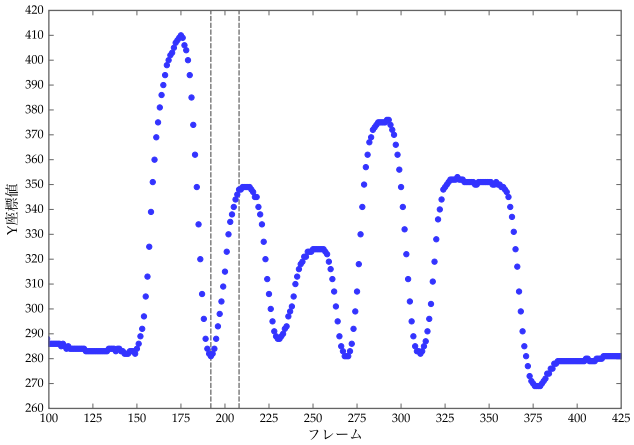

図 4　下唇の y 座標値の時系列変化

Fig. 4　Sequential value of the Y coordinate of lower lip.

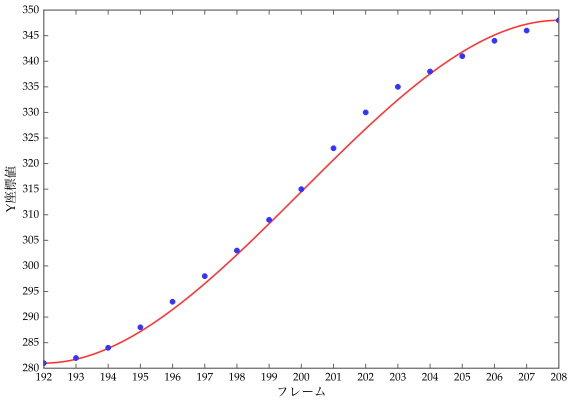

図 5　下唇が動く際の y 座標値の点と 3 次式の曲線

Fig. 5　Points of the Y coordinate value of lower lip and superimposed curve of the cubic equation.

まず，発話語句を仮名表記（平仮名または片仮名）で入力する．そして，入力語句から “口形順序コード” と呼ばれる記号列を生成する [6]．この口形順序コードは，入力された語句を発話する際に形成される口形の順序を表現しているため，CG 技術のモーフィングを用いて口形変形画像を生成し，発話映像を生成する [7]．モーフィングを行う場合，**表 1** に示す基本口形画像をキーフレームとして口形変形画像を生成するが．ただし，口形画像には唇領域と歯領域，口内領域が存在しており，口形によってその有無が異なるため，それぞれの領域の変形を考慮する必要がある．例えば，ア口形（*A*）からエ口形（*E*）へ変形する場合は，それぞれ歯領域と口内領域が存在するため対応する領域間で変形を行えば良いが，イ口形（*I*）からオ口形（*O*）へ変形する場合は，歯領域と口内領域の有無が異なるため，対応する領域間での変形が行えない．そこで，存在しない領域は黒に近い暗い画像領域を想定して領域間のブレンドを行うこととする．

生成する発話映像を 60fps で表示する場合，2.1 で使用した発話映像から換算すると，口形の変形に要するフレーム

表 1　基本口形画像と各画像内の歯領域と口内領域の有無
Table 1　Images of basic mouth shape. The teeth area and buccal area in the images.

基本口形	基本口形画像	歯領域	口内領域
A		有	有
I		有	無
U		有	無
E		有	有
O		無	有
X		無	無

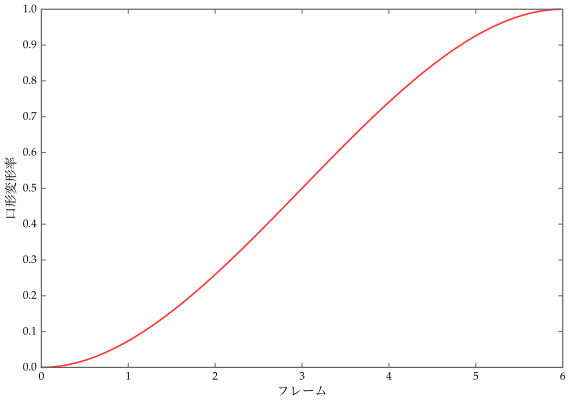

図 6　口形変形画像生成時の口形変形率
Fig. 6　Deformation rate for generation of deformed mouth shape image.

数は 7 となる．そこで，**図 6** に示す口形変形フレームの口形変形率から口形変形画像を生成する．実際には，最初と最後のフレームは基本口形画像となるため，生成する口形変形画像は 5 枚となる．例として，生成した口形変形画像を**図 7** に示す．図 7 では閉唇口形からア口形への変形過程を示しており，フレーム 0 とフレーム 6 はそれぞれ変形前後の基本口形画像でこれらがキーフレームとなっている．

4.　実験

提案方法を用いた発話映像の生成実験を実施した．実験は 2 種類行い，第 1 の実験（以降，実験 1 とする）では発話映像の生成に関する評価を行い，第 2 の実験（以降，実験 2 とする）では生成された発話映像に関する評価を行った．以下に，それぞれの実験について結果を示す．

表 2　画像生成を行ったコンピュータ環境
Table 2　Computer and development environment that were used for generation of deformed mouth shape images.

CPU	3.4GHz Intel Core i7
メモリ	32GB
OS	Mac OS X 10.8.3
プログラミング言語	C++
コンパイラ	Intel C++ Composer XE 2013 Update 3
画像処理ライブラリ	OpenCV 2.4.5
キーフレーム画像サイズ	640×480 ピクセル

4.1　実験 1

実験 1 では，モーフィングを用いた口形変形画像の生成にどのくらいの時間がかかるかを計測した．画像生成に使用したコンピュータやプログラム開発環境を**表 2** に示す．その結果，画像 1 枚あたりの生成に平均 293 ミリ秒かかったが，発話映像を 60fps のフレームレートで表示するならば 16 ミリ秒以下で生成する必要がある．従って，現在のコンピュータ環境ではリアルタイムにモーフィングで口形を変形させていくことは困難であることがわかった．

そこで，発話映像生成時に必要となる全ての口形変形画像を事前に生成しておき，発話映像生成時にはその中から必要な画像を組み合わせて発話映像を生成することとした．発話映像を生成する場合，終口形から初口形と初口形から終口形，終口形から終口形への変形があり，それぞれの変形に要するフレーム数が異なる．そのため，各口形の組につき，31 枚の口形変形画像を生成した．つまり，ア口形からイ口形への口形変形画像を 31 枚，同様に他の口形の組についても 31 枚の画像を生成し，合計で 465 枚の口形変形画像を生成した．これにさらに，基本口形の画像 6 枚を合わせた 471 枚の画像で発話映像を生成した．その結果，発話映像を 60fps で表示させることが可能となった．現在のスマートフォンはパソコンに比べてさらに性能が低いため，この方法は将来的にスマートフォン向けアプリケーションを開発する場合にも有効であると考える．一方，あらかじめ生成しておく画像量が多いためデータ容量が大きくなるが，スマートフォンの CPU 性能に比べてメモリ容量は比較的容易に増やすことができるため，この問題も解決できると考える．

4.2　実験 2

実験 2 では，実験 1 で生成した発話映像に関する評価を行った．評価は，被験者の主観による評価と発話語句を推測する評価を行った．発話速度を “普通” と “遅い”，“速い” の 3 段階に設定して発話映像を生成し，その滑らかさや自然さを評価した．評価を行った被験者は，読唇術の経

図 7 生成した口形変形画像（閉唇口形からア口形への変形）

Fig. 7 Generated mouth shape images when closed mouth shape changes into mouth shape /a/.

表 3 実験 1 の発話語句とその口形順序コード

Table 3 Utterance words and their mouth shape sequence code for the first experiment.

#	発話語句	口形順序コード
1	カタツムリ	-AIA-UXU-I
2	川下り	-AUA-UIA-I
3	紙芝居	-AXI-IXA-I
4	アセスメント	-AIE-UXE-IUO
5	スポットライト	-UXO-UUOIA-IUO

表 4 発話映像の主観評価結果

Table 4 Result of mean opinion score of utterance images.

#	発話語句	普通	遅い	速い
1	カタツムリ	3.6	4.2	3.1
2	川下り	3.9	3.2	3.0
3	紙芝居	2.3	2.6	2.5
4	アセスメント	4.1	4.1	3.8
5	スポットライト	3.4	3.5	2.7
平均		3.5	3.5	3.0

験のない神奈川工科大学情報工学科に在籍する日本人の 4 年生 15 名である．

4.2.1 発話映像の主観評価

この実験では，**表 3** に示す 5 つの発話語句について，それぞれの発話速度の映像を 3 回ずつ示した．各語句のそれぞれの発話速度の映像に対して，自然に感じられれば 5，不自然に感じられれば 1 の 5 段階評価をしてもらった．評価の平均値を**表 4** に，その評価を棒グラフに表したものを**図 8** に示す．

この結果，“アセスメント” に関しては，それぞれの発話速度で 4 に近い評価が得られたため，自然な発話映像が生成できたと言える．しかしながら，“紙芝居” については各発話速度で 3 を下回る評価となったため，不自然な発話映像となったと言える．この要因は，連続する “み” と “し” にあると考える．“み” と “し” は同じイ段の音であるため，2 つの音での口形の変化はない．そのため，被験者からのコメントにも，“み” と “し” の区別がつかなかったという意見や，“みー” のように長音として感じられたという意見があった．ただし，これは，実際の発話においても口形の変化がないため，被験者に読唇術の経験がないためか，口唇周辺のみの映像だったためと考えられる．

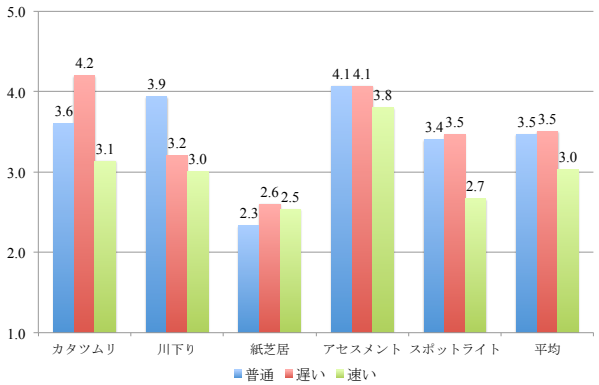

図 8 発話映像の主観評価グラフ

Fig. 8 Bar chart of the result of mean opinion score.

表 5 実験 2 の発話語句とその口形順序コード

Table 5 Utterance words and their mouth shape sequence code for the second experiment.

#	発話語句	口形順序コード
1	スイカ	-U-I-A
2	カラス	-AIA-U
3	筆箱	-UIEXA-O
4	川崎	-AUAIA-I
5	腕時計	-UIEUO-E-I
6	裁判所	IA-IXA-IUO

全体平均は，各発話速度で 3 を少し超えた評価であるため，自然とも不自然とも言えない結果となった．

4.2.2 発話語句の推測

この実験では，**表 5** に示す語句の発話映像を生成し，発話速度を “普通” と “遅い” に設定した映像を 3 回ずつ示した．その際，被験者には発話語句を告げずに，語句 A〜F として発話映像を示した．ただし，読唇経験を持たない被験者にとって，単純に発話映像から語句を推定することは困難であると考えたため，**表 6** の分類に示す内容をヒントとして提示した．

各語句の正解率を**表 7** に，そのグラフを**図 9** に示す．語句 B（カラス）や語句 C（筆箱）に関しては高い正解率が得られたが，語句 F（裁判所）に関しては 1 名の被験者しか正解しなかった．語句 F では，最後の “しょ” の部分は分かったが，その前の部分が分からなかったというコメントがあった．また，この語句は他の語句に比べて初口形 [4] と

表 6　各語句の発話内容と被験者へ提示した分類
Table 6　Categories of the utterance word shown to subjects.

#	提示語句	発話語句	分類
1	語句 A	スイカ	野菜または果物
2	語句 B	カラス	鳥
3	語句 C	筆箱	文具
4	語句 D	川崎	都市名
5	語句 E	腕時計	身につけるもの
6	語句 F	裁判所	建物または場所

表 7　発話語句の正解率
Table 7　Correct answer rates of utterance images.

#	語句	普通	遅い
1	語句 A（スイカ）	60.0%	73.3%
2	語句 B（カラス）	80.0%	86.7%
3	語句 C（筆箱）	80.0%	86.7%
4	語句 D（川崎）	66.7%	80.0%
5	語句 E（腕時計）	53.3%	73.0%
6	語句 F（裁判所）	6.7%	6.7%
平均		57.8%	67.8%

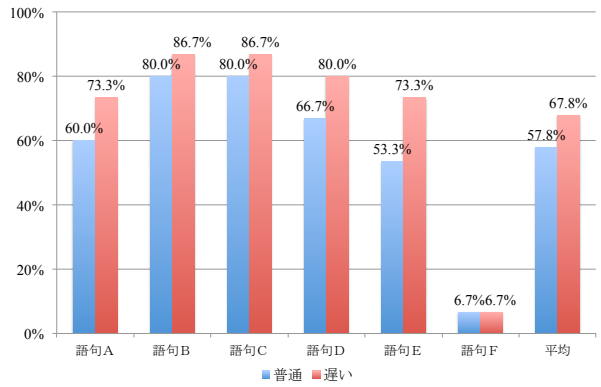

図 9　発話内容の正解率グラフ
Fig. 9　Bar chart of the correct answer rates.

呼ばれる口形が多かったために，語句の推測が困難になったと考える．

しかしながら，被験者が読唇の経験がないことを踏まえると，語句 F を除けば高い正答率が得られたのではないかと考える．このことから，実験 1 では，生成した発話映像が自然であるという評価は得られなかったが，発話内容を理解できる映像が生成できていると見なすことができる．また，ほとんどの語句において，遅く発話した映像の方が正答率が高くなっており，平均でも 10%正答率があがっているため，発話速度による効果も得られたと考えられる．

5. まとめ

本論文では，聴覚障害者が読唇術を習得するためのトレーニング用教材に向けた，日本語語句の発話映像自動生成について提案した．高いフレームレートで撮影できるカメラを用いて口形の変化を分析し，そこから口形の変化率を導出した．この口形変化率を基に生成した発話映像の実験からは自然であるという評価は得られなかったが，発話映像から発話語句を推測する実験では高い正解率が得られたため，正しい口形変化が表現できたと言える．また，発話速度を下げることによって発話語句の正答率が向上したため，発話速度を変化させることの有効性を示すことができたと考える．今後は，いくつかの違和感が残る点を修正しながら，Android 端末等のトレーニングアプリケーションの構築へ反映させていきたいと考えている．

謝辞　本研究は JSPS 科研費 23700672 の助成を受けたものです．

参考文献

[1] Asa, H. and Bertil, L.: Visual Speech Synthesis with Concatenative Speech, *Auditory-Visual Speech Processing*, pp. 181–184 (1998).
[2] 垣原清次, 中村哲, 鹿野清宏：HMM を用いた自然な発話動画像合成, 電子情報通信学会論文誌 D-II, Vol. J83-D-II, No. 11, pp. 2498–2506 (2000).
[3] 川本真一, 下平博, 新田恒雄, 西本卓也, 中村哲, 伊藤克亘, 森島繁生, 四倉達夫, 甲斐充彦, 李晃伸, 山下洋一, 小林隆夫, 徳田恵一, 広瀬啓吉, 峯松信明, 山田篤, 伝康晴, 宇津呂武仁, 嵯峨山茂樹：カスタマイズ性を考慮した擬人化音声対話エージェントツールキットの設計, 情報処理学会論文誌, Vol. 43, No. 7, pp. 2249–2263 (2002).
[4] 宮崎剛, 中島豊四郎：日本語発話時の特徴的口形のコード化と口形変化情報表示方法の提案, 電気学会論文誌 C, Vol. 129, No. 12, pp. 2108–2114 (2009).
[5] Tsuyoshi, M., Toyoshiro, N. and Naohiro, I.: Evaluation for an Automatic Generation of Lips Movement Images Based on Mouth Shapes Sequence Code in Japanese Pronunciation, *Proc. of Japan-Cambodia Joint Symposium on Information Systems and Communication Technology, 2011 (JCAICT 2011)*, pp. 89–92 (2011).
[6] 宮崎剛, 中島豊四郎：スマートフォン向け読唇トレーニングアプリケーションの試作と評価, マルチメディア, 分散, 協調とモバイル (DICOMO2012) シンポジウム, pp. 1863–1868 (2012).
[7] 宮崎剛, 中島豊四郎：口形順序コードを用いた発話映像自動生成方法, 第 9 回情報科学技術フォーラム (FIT2010) 講演論文集, 第 3 分冊, pp. 671–672 (2010).